Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training

Vardges Ter-Hovhannisyan,[1,4] Alexandre Lomsadze,[2,4] Yury O. Chernoff,[1] and Mark Borodovsky[2,3,5]

[1] *School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332, USA;* [2] *Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia 30332, USA;* [3] *Computational Science and Engineering Division, College of Computing, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*

We describe a new ab initio algorithm, GeneMark-ES version 2, that identifies protein-coding genes in fungal genomes. The algorithm does not require a predetermined training set to estimate parameters of the underlying hidden Markov model (HMM). Instead, the anonymous genomic sequence in question is used as an input for iterative unsupervised training. The algorithm extends our previously developed method tested on genomes of *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. To better reflect features of fungal gene organization, we enhanced the intron submodel to accommodate sequences with and without branch point sites. This design enables the algorithm to work equally well for species with the kinds of variations in splicing mechanisms seen in the fungal phyla Ascomycota, Basidiomycota, and Zygomycota. Upon self-training, the intron submodel switches on in several steps to reach its full complexity. We demonstrate that the algorithm accuracy, both at the exon and the whole gene level, is favorably compared to the accuracy of gene finders that employ supervised training. Application of the new method to known fungal genomes indicates substantial improvement over existing annotations. By eliminating the effort necessary to build comprehensive training sets, the new algorithm can streamline and accelerate the process of annotation in a large number of fungal genome sequencing projects.

[Supplemental material is available online at www.genome.org. The new software program GeneMark-ES version 2 is freely available for download from http://exon.gatech.edu/genemark/gmhmm-es-2008.]

Reliable ab initio gene prediction in eukaryotic genomic sequences remains an open problem in spite of impressive progress made in developing gene prediction algorithms (Burge and Karlin 1997; Krogh 1997; Parra et al. 2000; Reese et al. 2000; Stanke and Waack 2003; Guigo et al. 2006). Much attention has been given to developing alternative, extrinsic methods that use EST/cDNA to genome mapping, spliced alignments of known protein sequences, or patterns of conservation between related genomes (Gelfand et al. 1996; Mott 1997; Mathe et al. 2002; Birney et al. 2004; Stanke et al. 2008). The extrinsic methods exhibit, as a rule, high specificity (Sp), while ab initio, intrinsic methods show high sensitivity (Sn). These properties make methods of both types indispensable in genome annotation pipelines. For accurate statistical description of protein-coding regions, efficient ab initio algorithms employ the fifth-order three periodic Markov chain models incorporated into hidden Markov models (HMMs) (Kulp et al. 1996; Burge and Karlin 1997) The number of algorithm parameters, several thousands, is high; thus, a training set of ~1000 experimentally validated genes is necessary for parameter estimation. Compilation of such large training sets represents a bottleneck in genome annotation pipelines and a practical challenge that hampers the use of ab initio gene prediction algorithms.

To circumvent this difficulty, we developed a gene finder able to extract model parameters from the original genomic sequence (Lomsadze et al. 2005). We demonstrated that unsupervised model training, well known in prokaryotic gene finding (Audic and Claverie 1998; Hayes and Borodovsky 1998; Salzberg et al. 1998; Besemer et al. 2001; Larsen and Krogh 2003; Delcher et al. 2007), is also feasible for eukaryotes. Particularly, for genomes of *Arabidopsis thaliana*, *Drosophia melanogaster*, and *Caenorhabditis elegans*, the accuracy of the gene finder with unsupervised parameter estimation matched the accuracy of a conventional supervised gene finder. Nevertheless, variability in eukaryotic gene organization is by far greater than that has been observed in prokaryotes; thus, automatic algorithms specialized for genomes that share some important features may provide higher accuracy. Along this line, we have explored the opportunity to better analyze genomes where a significant part of the information for intron splicing is carried by a branch point (BP) site. Many fungal genomes fall into this category. Fungal genomes are relatively densely populated with genes (with protein-coding regions occupying from 27.8% of genome in *Neurospora crassa* to 52.6% in *Botritus cineria*), and these genes exhibit significant variation in exon-intron structure (Fig. 1).

The BP site models with parameters estimated by supervised training were used earlier in several gene finding algorithms (Lukashin et al. 1992; Hebsgaard et al. 1996; Burge and Karlin 1997; Stanke and Waack 2003). Structures of the BP models used varied from a positional frequency matrix in NetGene2 (Hebsgaard et al. 1996), IntronScan (Lim and Burge 2001), and GipsyGene (Neverov et al. 2003) to the second-order windowed weight array matrix (WWAM) generating a 21-nt-long sequence in GenScan (Burge and Karlin 1997), to the third-order WWAM generating a 32-nt-long sequence in AUGUSTUS (Stanke and Waack 2003). Derivation of a positional frequency matrix proceeded

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail borodovsky@gatech.edu; fax (404) 894-4243.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.081612.108.

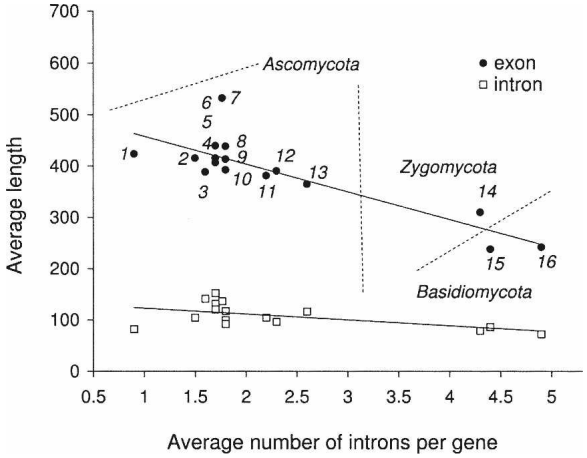

Figure 1. Annotations of the 16 fungal genomes demonstrate significant variability in gene organization: both in the average size of exons and introns as well as in the number of introns per gene (1, *Schizosaccharomyces pombe*; 2, *Aspergillus niger*; 3, *Botrytis cinerea*; 4, *Stagonospora nodorum*; 5, *Fusarium oxysporum*; 6, *Magnaporthe grisea*; 7, *Neurospora crassa*; 8, *Fusarium graminearum*; 9, *Fusarium verticillioides*; 10, *Sclerotinia sclerotiorum*; 11, *Aspergillus terreus*; 12, *Aspergillus nidulans*; 13, *Coccidioides immitis*; 14, *Rhizopus oryzae*; 15, *Coprinus cinereus* (also known as *Coprinopsis cinerea*); 16, *Cryptococcus neoformans*).

simultaneously with the putative BP sites alignment constructed by simulated annealing (Lukashin et al. 1992), an HMM-based EM algorithm (Hebsgaard et al. 1996), or a Markov chain Monte Carlo (MCMC) method (Thompson et al. 2003).

A fundamental question, "how much information is required by the recognition mechanism that controls splicing of short introns to accurately identify the locations of introns in primary transcript sequences and how this information is distributed?", was addressed by computational modeling (Lim and Burge 2001). The analysis of five diverse species demonstrated that in *D. melanogaster*, *A. thaliana*, *C. elegans*, and *Homo sapiens*, the BP site contributes less than 5% of information necessary for accurate intron splicing; however, in yeast (*Saccharromyces cerevisiae*), a species with just ~200 spliced genes, the BP site contributes ~40% of information. More recently, a study of five genomes from phyla Ascomycota and Basidiomicota has also shown a presence of strongly conserved BP sites in fungal species with genes interrupted by introns much more frequently than genes in yeast genomes (Kupfer et al. 2004).

Existing evidence of a significant role of the fungal BP sites suggests that the HMM model underlying an ab initio algorithm should account for introns possessing conserved BP sites. To address this issue, we introduced an enhanced intron submodel that accommodates intron sequences with and without BP sites. In the context of this project, estimation of parameters of this submodel has to be done by unsupervised training. To solve the training task, we combined two unsupervised training approaches: the Viterbi training (Durbin et al. 1998) and the MCMC method (Thompson et al. 2003). The MCMC algorithm is run in full at each iteration of the Viterbi training procedure to generate an update of parameters of the new intron submodel. Since an HMM fit to data is largely determined by the values of emission probabilities (Mitrophanov et al. 2005), the emission probabilities are updated in each iteration while the probabilities of transition between hidden states are updated once in a few iterations.

To test the algorithm, we selected 16 genomes from the three fungal phyla (Supplemental Fig. S1) and compiled sets of sequences containing EST validated genes. We demonstrated that the self-training algorithm with the enhanced intron submodel predicts fungal genes with higher accuracy than either the original self-training algorithm or algorithms with supervised training. Application of the new algorithm indicated that a few of the 16 selected genomes are likely to be overannotated while several are underannotated. A number of fungal genome sequencing projects (more than 300 projects have been registered at www.genomesonline.org) may benefit from the new method higher accuracy, as well as from the convenience of an immediate transition from sequencing to the gene prediction and annotation stage without taking time on the tedious extraction of a training set of manually curated genes.

Methods

Sequence data

Genomic and EST sequences of 16 fungal genomes (Table 1; Supplemental Tables S1, S2) are from GenBank (www.ncbi.nlm.nih.gov/Genbank) and the Broad Institute (www.broad.mit.edu). The fungal genomes selected for this study varied in size from 20 Mb (*Cryptococcus neoformans*) to 60 Mb (*Fusarium oxysporum*) and in GC content from 36% (*Rhizopus oryzae*) to 52% (*Aspergillus terreus*). These genomes (as annotated) show a significant variation in gene number and organization of exon-intron structure: from 5055 genes in *Schizosaccharomyces pombe* to 17,735 genes in *F. oxyporum*; an average number of introns per annotated gene varied from 0.9 (*S. pombe*) to 4.9 (*C. neoformans*) (Table 1).

Iterative unsupervised estimation of model parameters used in the original algorithm

A conventional ab initio gene prediction algorithm with assigned HMM parameters produces a sequence parse into protein-coding and noncoding regions (Burge and Karlin 1997). Conversely, in an iterative self-training algorithm, a given sequence parse is used for HMM parameters re-estimation (Lomsadze et al. 2005). To obtain an initial sequence parse, our original self-training algorithm uses (1) three-periodic Markov chain models with parameters inferred from codon frequencies determined as functions of genome GC content (Besemer and Borodovsky 1999), (2) zero-order model of noncoding region initialized with genome-specific nucleotide frequencies, and (3) minimal size (2 nucleotide [nt]) models of canonic splice sites. Understandably, sequence parses made at the start of the process are treated with caution. Only those exon-intron structures that produce open reading frames (ORFs) with sufficient length and coding potential are selected into an update of a training set (Lomsadze et al. 2005). Further consecutive steps (rounds) of sequence parsing, selection of a training set and parameter re-estimation continue, as supposed by the logic of the Viterbi training (Durbin et al. 1998), with sequence parse k producing the kth version of the parameter set. In contrast with the conventional Viterbi training, the HMM architecture is allowed to change in iterations; e.g., the model of acceptor site grows from its initial simplistic form of just two canonical letters AG to the first-order Markov model generating 21-nt-long sequence fragments. The iterative process comes to conversion at a point where the predicted sequence

Table 1. Characteristics of the 16 fungal genomes and of the complements of predicted and annotated genes

Species	Estimated genome size (Mb)	GC content (%)	No. of genes		No. of single exon genes		No. of introns per gene		No. of introns per spliced gene	
			Annotated	Predicted	Annotated	Predicted	Annotated	Predicted	Annotated	Predicted
A. nidulans	31	50	10,701	10,445	1453	2278	2.3	2.0	2.7	2.6
A. niger	34	50	14,101	11,342	1538	2405	1.5	2.1	1.7	2.7
A. terreus	29	52	10,406	10,859	1538	2288	2.2	2.1	2.6	2.7
B. cinerea	26	43	16,448	11,890	4316	2624	1.6	1.8	2.2	2.3
C. immitis	29	46	10,457	8435	1449	1903	2.5	2.0	2.9	2.6
C. cinereus	38	51	13,544	12,952	1011	1480	4.4	4.5	4.8	5.1
C. neoformans	20	48	7302	7246	252	441	4.9	4.8	5.1	5.1
F. graminearum	40	48	13,332	12,426	3096	3126	1.8	1.7	2.3	2.3
F. oxysporum	60	48	17,735	20,843	4409	6222	1.7	1.6	2.3	2.3
F. verticillioides	42	48	14,179	14,716	3536	3922	1.8	1.7	2.4	2.4
M. grisea	40	51	12,841	11,850	3000	2916	1.7	1.6	2.2	2.1
N. crassa	39	49	9826	9679	1832	2304	1.8	1.5	2.2	1.9
R. oryzae	40	36	17,467	16,477	3413	3962	2.3	3.0	3.8	3.5
S. pombe	12	36	5055	4,913	2764	2616	0.9	1.0	2.0	2.2
S. sclerotiorum	39	51	14,522	11,119	3278	2490	1.8	1.8	2.3	2.4
S. nodorum	37	51	16,597	13,707	2359	3582	1.7	1.6	2.0	2.2

Gene predictions were generated by the algorithm with enhanced intron submodel at the convergence point of self-training. Annotation data from EMBL (*A. niger*), NCBI (*S. pombe*), and Broad Institute (http://www.broad.mit.edu/) as of May 2008.

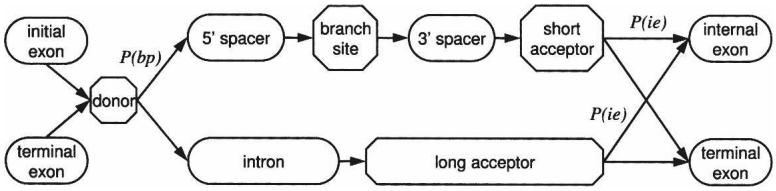

Figure 2. The hidden state diagram for the enhanced intron submodel (direct strand).

parse $k + 1$ is not distinguishable from the previous parse k. Typically, conversion of the original algorithm was observed by the sixth iteration.

Introduction of the enhanced intron submodel

The new intron submodel provides two alternative paths of hidden states for an intron sequence (Fig. 2). The bottom path, the one that existed in the original algorithm, consists of "intron" and "long acceptor" states. The new top path has four hidden states: "upstream spacer," "BP site," "downstream spacer," and "short acceptor." The "upstream spacer" generates a nucleotide sequence situated between the donor site and the BP site; nucleotide composition of this sequence and the sequence generated by the "intron" state in the bottom path are the same. This setting is justified by earlier reports (Kupfer et al. 2004), as well as by our observations. The "downstream spacer" generates a sequence between the BP site and the acceptor site modeled by the first-order homogeneous Markov model. Nucleotide composition of downstream spacers is reduced in guanine and shows strong asymmetry between frequencies of adenine and thymine, with preference for thymine. The length distribution of the downstream spacer sequences has a skewed bell shape (Supplemental Fig. S2) with a clear lower bound that is apparently maintained by negative selection. The "short acceptor" state that models just 2 nt upstream of AG (compared with 18 nt modeled by the "long acceptor") is appropriate for fungal genomes whose introns do not possess a poly-pyrimidine (poly-Y) tail upstream of the acceptor site while having a strongly conserved BP site. However, introns in some fungal genomes do have the poly-Y tail (see below). The probability of transition from a donor state to a BP-holding intron (top path), *P(bp)*, is genome specific. Another genome-specific transition probability in the intron submodel is the probability of transition, *P(ie)*, to internal exon (Fig. 2).

Changes in the procedure of parameter estimation

The iterative self-training procedure follows the path of the original algorithm up to the fourth iteration (Lomsadze et al. 2005). During the initial iterations, the probability value *P(bp)* in the intron submodel (Fig. 2) was set to zero. Parameterization of the new intron submodel begins with assignment of a nonzero *P(bp)* value, 0.5, at the fifth iteration. Unsupervised estimation of the parameters of the top path of the enhanced intron submodel focuses on derivation of parameters of the BP model. A model of the BP site per se is a positional Markov model. An alignment of true BP sites, should they be known, provides positional frequency statistics of nucleotides or dinucleotides for the BP model parameter estimation. We use the MCMC method, the Gibbs sampler algorithm (Lawrence et al. 1993; Thompson et al. 2003), to construct a multiple alignment of intron sequences scored within the "target windows" that are used to reveal evolutionarily conserved BP sites (the details are given in the next section). A "target window" positioned in a given intron creates a parse delineating two more intron sections, the BP upstream and the BP downstream sequences. The set of upstream (downstream) sequences is used to define parameters of the compositional model and the length distribution of the upstream (downstream) spacer. Graphic illustration of several components of the intron model computed for *Stagonospora nodorum* is given in Figure 3. After completion of the fifth iteration, the value of *P(bp)* is changed from 0.5 to a ratio of the number of introns predicted (emitted) via the hidden states of the "top path" to the total number of introns (Supplemental Table S3). Notably, the value of this ratio correlates with the ratio of introns with weak BP sites, as identified by the Gibbs sampler alignment algorithm, to the total number of introns in the Gibbs sampler alignment (data not shown).

The value of the probability of a transition from an intron to an exon state, *P(ie)*, relates to the average number of exons per gene. It is assumed that empirical distribution of the number of exons per gene follows the shape of geometric distribution (Supplemental Fig. S3). Therefore, the value of *P(ie)* is estimated by fitting a geometric distribution to the distribution of number of exons per gene that emerged as a result of gene prediction. In the computations for a given genome, we initially set *P(ie)* to 0.5 and do not change this value through the iterations until convergence. Then the corrected value, *P(ie)**, is estimated from the updated distribution of number of exons per gene as predicted. Subsequently, the training procedure is repeated until convergence with the new fixed value *P(ie)** and *P(ie)* is re-estimated. If the updated estimate, *P(ie)***, is close to *P(ie)**, which is usually a case, we accept *P(ie)** as a final value of *P(ie)*.

Interestingly, the noticeable differences between the shapes of empirical distributions of a number of exons per gene (Supplemental Fig. S3) and their theoretically expected shapes (the straight lines) call for a revision of the HMM architecture used in the algorithm, a subject of a separate study.

Estimation of parameters of the new intron submodel

In running the Gibbs sampler, we did not attempt to streamline the sampling process by fixing the BP nucleotide "A" (Hebsgaard et al. 1996) by masking a part of a sequence to reduce the search space (Lim and Burge 2001) or by creating initial profiles based on a knowledge derived from related species (Neverov et al.

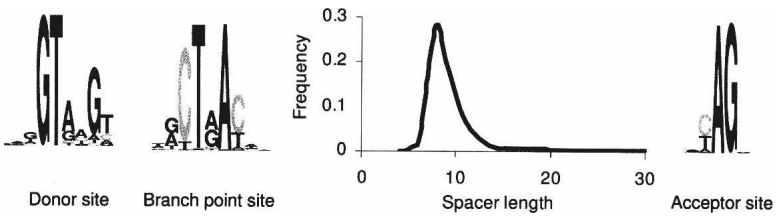

Figure 3. The logos of splice sites and branch point models as well as the graph of the length distribution of the spacer between branch point and acceptor site (as determined for the *S. nodorum* genome).

2003). Still, we selected a subset of all predicted introns with lengths close to the one observed with maximum frequency (±10 nt); additionally, we reduced each intron sequence to a 50-nt fragment situated upstream of the acceptor site (if the sequence is longer than 50 nt). The rationale for doing this is as follows. First, very short or very long introns are likely to be erroneous; also, even in truly long introns, the BP site may be situated further away than 50 nt from the acceptor site. Second, reducing fragment lengths to 50 nt makes the Gibbs sampler procedure more efficient. The length of the positional frequency model of the BP site (the "target window" length) was chosen to be 9 nt based on computational experiments with window lengths from 7–11 nt (data not shown). The Gibbs sampler was run in the site sampler mode assuming a single motif per sequence (http://bayesweb.wadsworth.org/web_help_text.Gibbs_versions.html). Given that a single iteration of the Gibbs sampler includes repositioning of each sequence involved in alignment, the number of iterations used was 500 (the default setting). At the final step, the positional frequency model was derived from only those motifs in a given intron sequence that appeared within the target window more than 50% of the time at the stage of the near optimal sampling, normally reached after the first 250 iterations. Since we did not require the presence of adenine "A" in the branching position, as it is not a necessity in nature, nucleotides T, C, or G could appear in the branching position with small frequencies. To more accurately estimate these small values, we have reduced the Gibbs sampler default 10% fraction of pseudo counts to 0.1%. Next, we used the model of BP sites derived from a subset of introns to determine the most likely BP positions in the remaining introns as positions possessing the highest score with regard to the BP model. All predicted BP positions were then used to delineate more precisely the length distribution of upstream spacers. Also, we used the set of downstream spacer sequences as a training set for the compositional model of downstream spacer sequence. Finally, with parameters of the enhanced intron submodel in place, the algorithm is ready for the next iteration.

Preparation of test sets

Expressed sequence tags (ESTs) data were available for each of the 16 fungal species (Supplemental Table S2). For assessment of the algorithm performance, we used EST validated genes identified by a local pipeline mapping EST to genomic DNA (A. Kislyuk, A. Lomsadze, and M. Borodovsky, unpubl.). We have compiled test sets of two types. A test set of type I included only complete genes validated by EST sequences; a test set of type II included both complete and incomplete genes (Supplemental Table S4).

The EST to genome mapping proceeded as follows. First, genome-specific ESTs were aligned to genomic sequence using the program BLAT (Kent 2002). If (1) an individual alignment failed to cover more than 90% of the EST sequence or (2) any of the exons in alignment had an EST to genomic sequence identity below 90%, the whole EST sequence was discarded. Alignments that delineated introns with noncanonical splice sites were discarded as well. *Integrated transcripts* (alignment clusters) were generated from the retained EST to genome alignments by joining alignments that shared the same introns. If members of integrated transcripts obtained in this way disagreed on some other introns, the alignment cluster was split into noncontradictory transcripts of maximal length. At the next step, each transcript was analyzed for the presence of a noninterrupted protein-coding region in the GeneMark.hmm-P program (Lukashin and Borodovsky 1998) modified to use the Kozak context model along with the heuristic model of protein-coding regions (Besemer and Borodovsky 1999), and predicts protein-coding regions as whole or partial ORFs. Normally, just one transcript from an alignment cluster would pass the filter for the presence of only one gene. Otherwise, if two or more distinct transcripts with conflicting gene structures would appear, we assumed that the locus in question is hosting a gene with alternative splicing; then the whole integrated transcript was removed from consideration. Subsequent mapping of a transcript to genomic DNA defined either a complete (containing both 5′ and 3′ UTR, with translation start and end predicted by the modified GeneMark.hmm-P program) or incomplete gene. The rules of thus described EST-to-DNA mapping as well as transcript selection are rather stringent. As a consequence, the test sets of complete genes (type I test sets) turned out to be larger than 150 genes for only four of the fungal species analyzed here: *Coprinus cinereus*, *Coccidioides immitis*, *Fusarium verticillioides*, and *Magnaporthe grisea* (Supplemental Table S4). Additionally, we generated a type I test set for *S. pombe*, the most studied genome, by using annotated complete genes with protein products exactly matching *S. pombe* proteins in the SWISS-PROT database (Watanabe and Harayama 2001). The type II test sets, the test sets that include incomplete genes, were compiled for the remaining 11 fungal genomes (Supplemental Table S4). We should emphasize that the translation initiation and termination sites in the test set of complete genes were determined computationally; thus, the values of Sn and Sp determined for these sites should be taken with caution.

To create yet another type of a test set, we have connected 1277 manually curated complete *S. pombe* genes by random sequences emulating noncoding regions of the *S. pombe* genome with 31% GC content. The "intergenic" sequences of a given "artificial chromosome" were chosen to have one and the same length *L*; thus, we created instances of the *S. pombe* artificial chromosome for several *L* values ranging from 50 to 6000 nt. All gene sequences were placed in the direct strand of a chromosome. Computational experiments with the reverse complement of an artificial chromosome produced the same results as ones reported for the direct sequence.

Results

Algorithm testing and application to several fungal genomes

We applied the new algorithm to genomic sequences of fungal species from the phyla Ascomycota, Basidiomycota, and Zygomycota (Table 1; Supplemental Fig. S1); the set of 16 species spans large evolutionally distances and exhibits significant variability in genome size, gene number, average number of introns per gene, as well as in the number of repetitive elements (Supplemental Table S5). The average sizes of exons and introns vary from 250 to 450 nt and from 90 to 160 nt, respectively (Fig. 1). Notably, the data show that Ascomycetes have fewer introns per gene and longer average intron and exon sizes than Basidiomycetes and Zygomycetes. The analysis of each genome went through the following steps. First, the program was run on unannotated genomic sequences and produced the HMM model parameters along with gene predictions. Second, we used the test set generated for a given genome (see Methods) to assess the prediction accuracy in terms of Sn and Sp. Third, we compared genes predicted in a whole genome with its current annotation

and analyzed observed differences. We also used the *S. pombe* artificial chromosomes (see Methods) to assess the frequency of gene merging and splitting, frequency of false-positive gene predictions in intergenic sequences, as well as an error rate in the *exact* gene prediction. The *S. pombe* genome in our set has the least number of introns per gene, with 55% of its genes possessing no introns at all.

Note that the yeast-like genomes, with 90%–95% of the intronless genes, belong to a gray area where the ab initio gene prediction programs developed for either prokaryotes or eukaryotes generate sufficiently accurate predictions for the intronless protein-coding ORFs. An accurate prediction of the remaining multiple exon genes presents a challenge. A prokaryotic gene finder will at best predict all exons in a gene as separate ORFs. A eukaryotic gene finder could potentially produce a much better result. However, the small total number of introns in a yeast-like genome presents a challenge for both supervised and unsupervised training, since it poses a natural restriction on the size of a training set for estimation of parameters of the intron submodel. We are aware of this problem and are working on the algorithm for accurate prediction of the intron containing genes in the yeast-like genomes.

Assessment of the accuracy of gene prediction

We demonstrated on test sets of complete and incomplete genes (see Methods) that transition to the enhanced intron submodel significantly improved Sn and Sp of prediction of splice sites, introns, and internal exons (Table 2; Supplemental Table S6). Notably, Sn was boosted higher than Sp in almost all categories (Table 2; Fig. 4; Supplemental Table S6). Still, for the *R. oryzae* genome, we observed just a slight increase in Sn and Sp values (Fig. 3). The likely reason behind this observation is as follows. Introns of the fungal species from phyla Ascomycota and Basidiomycota exhibit sequence conservation around the BP site while the conserved sequence near acceptor sites is short (and less informative). To the contrary, introns of *R. oryzae*, a member of phylum Zygomycota, possess a 20-nt-long poly-Y tail combined with a weak BP signal (Supplemental Figs. S4, S5). Weakly conserved BP sites and long poly-Y tails were observed also in introns of *Phycomyces blakesleeanuss*, a member of the same order of Zygomycota as *R. oryzae* (M. Bruce, A. Lomsadze, and M. Borodovsky, unpubl.). Notably, the estimated value of *P(bp)* for *R. oryzae* is equal to 0.22; thus the role of the BP model in intron prediction (and, arguably, its role in the splicing mechanism) is relatively small. Note that for the other 15 fungal species, we have *P(bp)* = 0.90, with the exception of *P(bp)* = 0.77 for *Magnaporthe grisea* (Supplemental Table S3). This distinction between members of the fungal phyla Ascomycota and Basidiomycota on one side and the phylum Zygomycota on the other indicates that the BP site centered type of the splicing mechanism (thought to be typical for fungi) or the poly-Y tail centered type (thought to be typical for higher eukaryotes) could evolve under selection pressure as alternative options in several eukaryotic lineages, including some lineages of fungal species.

Findings made upon derivation of the site models

We compared the models of BP and splice sites inferred by the MCMC method from intron sequences determined ab initio with the models derived by the same MCMC method from intron sequences mapped in genomic sequence via ESTs. The number of EST derived introns in a given genome varied from 1152 (*F. oxysporum*) to 7812 (*C. cinereus*); the set of introns identified by ab

Table 2. Accuracy of prediction of gene structure elements (Sn/Sp)

	S. pombe			*C. immitis*			*F. verticillioides*			*M. grisea*			*C. cinereus*		
	Intron model			Intron model			Intron model			Intron model			Intron model		
	Original	New	δ	Original	New	δ	Original	New	δ	Original	New	δ	Original	New	δ
Internal exon															
Sn	80.3	**88.2**	7.9	72.3	**82.8**	10.5	79.4	**85.6**	6.2	70.8	**89.2**	18.4	81.5	**85.0**	3.5
Sp	87.0	**89.6**	2.6	87.8	**93.0**	5.2	88.7	**91.2**	2.5	84.0	**91.7**	7.7	87.9	**89.7**	1.8
Intron															
Sn	84.0	**91.0**	7.0	75.4	**84.1**	8.7	85.6	**90.7**	5.1	76.6	**89.3**	12.7	84.7	**86.8**	2.1
Sp	89.1	**92.7**	3.6	85.6	**91.3**	5.7	91.5	**94.3**	2.8	84.2	**90.5**	6.3	89.2	**90.3**	1.1
Donor															
Sn	89.1	**93.1**	4.0	81.9	**87.0**	5.1	89.4	**92.2**	2.8	85.0	**92.1**	7.1	88.5	**89.6**	1.1
Sp	94.8	**95.1**	0.3	93.6	**94.4**	0.8	95.9	**96.2**	0.3	**94.4**	93.9	−0.5	**93.8**	93.6	−0.2
Acceptor															
Sn	86.1	**92.8**	6.7	77.7	**86.8**	9.1	87.1	**91.9**	4.8	79.7	**93.2**	13.5	86.1	**87.8**	1.7
Sp	91.3	**94.6**	3.3	88.3	**94.7**	6.4	93.3	**96.2**	2.9	87.6	**94.6**	7.0	91.2	**92.1**	0.9
Exon															
Sn	82.4	**88.0**	5.6	71.4	**79.7**	8.3	81.2	**85.3**	4.1	76.5	**88.0**	11.5	78.7	**81.2**	2.5
Sp	85.8	**89.2**	3.4	78.2	**84.6**	6.4	85.0	**87.9**	2.9	82.0	**89.1**	7.1	82.6	**84.3**	1.7
Initiation site															
Sn	85.8	**88.2**	2.4	75.9	**78.7**	2.8	81.0	**81.7**	0.7	84.6	**88.2**	3.6	**72.5**	**72.5**	0.0
Sp	86.4	**88.5**	2.1	76.8	**78.7**	1.9	81.3	**81.7**	0.4	86.1	**89.2**	3.1	**73.3**	72.9	−0.4
Termination site															
Sn	92.7	**94.2**	1.5	82.4	**86.1**	3.7	92.4	**94.8**	2.4	79.9	**89.3**	9.4	80.8	**83.2**	2.4
Sp	92.6	**94.2**	1.6	82.4	**87.1**	4.7	92.4	**95.4**	3.0	79.9	**89.3**	9.4	81.8	**84.8**	3.0
Nucleotide															
Sn	98.1	**98.6**	0.5	94.7	**96.1**	1.4	97.9	**98.8**	0.9	95.8	**98.2**	2.4	**95.8**	95.3	−0.5
Sp	99.4	**99.6**	0.2	95.3	**96.5**	1.2	96.5	**97.1**	0.6	93.5	**95.8**	2.3	94.6	**95.1**	0.5

Data are provided for the algorithm with the original and with the enhanced intron submodel. The Sn and Sp values were determined for the test sets of complete genes (test sets type I in Supplemental Table S4). Bold font shows the larger value out of the two adjacent ones. Differences in prediction accuracy are shown in the column labeled δ.

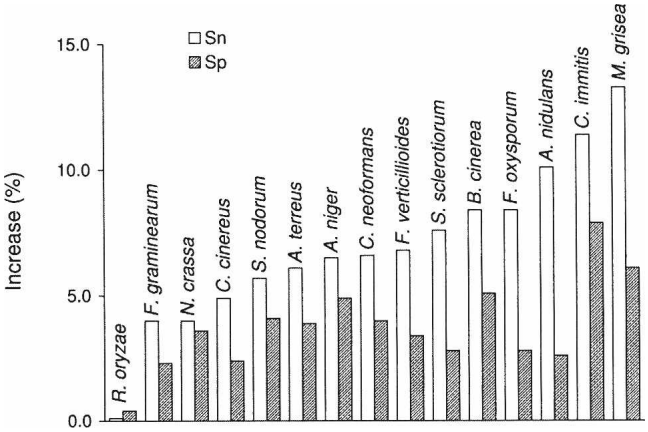

Figure 4. The increase in values of Sn and Sp of internal exon prediction achieved by the algorithm with enhanced intron submodel compared with the values of Sn and Sp of the algorithm with original intron submodel (at the points of algorithms' convergence).

initio algorithm in the same genome was observed to be three to 48 times larger. The relative entropies of the first-order models of BP and splice sites differed between two types of models by no more than 7% in almost all cases (Supplemental Table S7). As a rule, a BP site carries an amount of information comparable to what is carried by a donor site. Importantly, in all species but *R. oryzae*, the BP sites actually carry more information for accurate splicing than acceptor sites. In the predicted BP sites of all of the 16 fungal species except for *R. oryzae*, the consensus sequence of the positional nucleotide frequencies was CTNAC (Supplemental Fig. S5). The estimated frequencies of the canonical "A" in the BP position varied from 97% (*R. oryzae*) to 99% (*C. cinereus*). The most frequent length of the downstream spacer was 8–18 nt. We also observed characteristic unimodal shapes for the downstream spacer length distributions with the narrowest peak for *S. nodorum* (Supplemental Fig. S2). The value of the relative entropy for the spacer length distributions, with regard to the uniform distribution, can be used as a measure of "compactness" of the BP site localization. A low value of this parameter, in concert with a deviation of the consensus of the emerging BP model from CTNAC, usually pointed out to a problem with input sequence data; this could be due to either (1) a scarcity of evolutionarily conserved BP motifs in the sample (e.g., *R. oryzae*), or (2) significant enrichment with the extraneous sequences (due to erroneous acceptor site predictions).

Dynamics of the change in gene prediction accuracy upon algorithm iterations.

We traced the step-by-step changes in the accuracy of gene prediction by using the model parameters generated in subsequent iterations. We performed this analysis for the five test sets of complete genes (Set I in Supplemental Table S4). The values of Sn and Sp for predictions of exons, splice sites, initiation sites, and termination sites are shown in Supplemental Figure S6 as functions of iteration index. In the first iteration, one could expect to see a relatively low gene prediction accuracy of the algorithm employing heuristic models. However, this accuracy is high enough to allow for selecting a training set for reparameterization and improving the models to produce rather high Sp values in the second iteration: 40%–65% in internal exon prediction. Sp continues to grow in the third iteration where Sn remains in the range 20%–40%. However, in the next iteration, where length distributions of introns and exons are switched from uniform ones to estimated from predicted gene models, the Sn values jump to 58%–80%. Further improvement occurs in the fifth iteration, where we parameterize and start to use the enhanced intron submodel (Supplemental Fig. S6). We have to emphasize that the increases in Sn follow increases in Sp in the previous iteration. High Sp values indicate that the majority of current predictions are correct and ensure that the training set compiled for the next iteration includes a high percentage of correctly labeled sequences. Thus, not surprisingly, selection of a more reliable source for parameter estimation leads to better performance, particularly an increase in Sn.

Assessment of the accuracy of gene prediction in *S. pombe* artificial chromosomes

A test set having strictly one gene per sequence is not suitable for identification of errors in gene merging or for detection of incorrectly predicted genes in intergenic regions. However, for any given genome, it is difficult to find a large set of sequences possessing several validated genes in a row. Therefore, similar to an earlier implemented approach (Pavy et al. 1999), we used 1277 verified complete genes of *S. pombe* to construct a set of artificial chromosomes; "connectors" between genes were random sequences with fixed lengths ranging from 50 nt in the shortest chromosome to 6000 nt in the longest one (see Methods).

First, we observed that for all the lengths of intergenic regions the algorithm with the enhanced intron submodel had almost exactly the same performance as the original algorithm; both algorithms made a total of two to three splits in 1277 genes regardless of the length of the intergenic region. Second, we have detected that the new intron submodel makes the algorithm less prone to gene merging (Fig. 5). By taking into account the fact that 50% of intergenic regions in the *S. pombe* genome are longer than 750 nt (Supplemental Fig. S7) and that the real genes are located in both strands, which reduces the chance of merging, we estimate that the algorithm may merge ~20 out of 1000 (2%) of genes. Third, the new algorithm predicts false genes in intergenic regions at a slightly lower rate than the original algorithm (Supplemental Fig. S8). Thus, the rate of false-positive predictions is estimated at three per 1000 intergenic regions. Finally, in terms

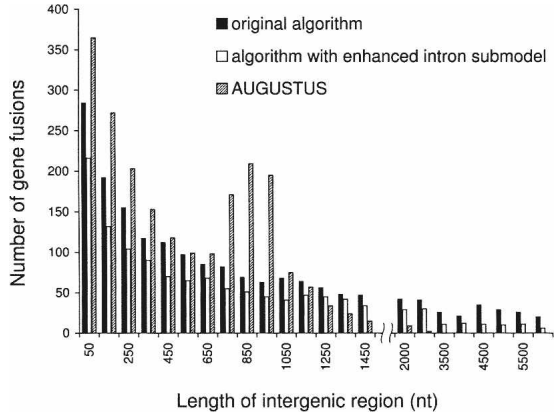

Figure 5. Statistics of erroneous gene fusions predicted in the *S. pombe* artificial chromosomes: the number of gene fusions per 1277 genes. The chromosomes differ in the length of random sequences emulating intergenic regions; all genes are placed in direct strand.

of exact gene prediction, the algorithm with the new intron submodel shows quite stable performance if intergenic regions are shorter than 1500 nt; here, it is able to exactly predict ~900 genes out of 1277 (Supplemental Fig. S9). The original algorithm makes exact predictions for ~800 genes in all types of the artificial chromosomes.

Comparison of the gene complements predicted and annotated in fungal genomes

We observed that predictions made by the new algorithm were in general agreement with the currently existing gene annotations both in terms of the total number of genes and the number of introns per spliced gene (Table 1). Particularly, in seven genomes out of 16—*Aspergillus nidulans*, *A. terreus*, *Coprinus cinereus*, *C. neoformans*, *F. verticillioides*, *N. crassa*, and *S. pombe*, the numbers of genes predicted and annotated were quite close to each other. The difference in total numbers of predicted and annotated genes was near or exceeded 1000 genes in genomes of nine species: *Aspergillus niger*, *Botrytis cinerea*, *C. immitis*, *Fusarium graminearum*, *F. oxysporum*, *M. grisea*, *R. oryzae*, *S. nodorum*, and *Sclerotinia sclerotiorum*. Only in three genomes, *A. terreus*, *F. oxysporum*, and *F. verticulloides*, was the number of predicted genes larger than the number of genes annotated (Table 1; Supplemental Fig. S10).

To further investigate this matter, we singled out in each of the nine genomes a set of "common" genes, both predicted and annotated, and then focused on the remaining two sets: genes annotated but not predicted and genes predicted but not annotated. The protein products of genes in both sets were used as queries for similarity searches against the "nr" protein sequence database (www.ncbi.nlm.nih.gov/) and the Conserved Domain Database (CDD) (www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml; Marchler-Bauer et al. 2007).

Genes annotated but not predicted

A rather small fraction of protein products of these genes had a statistically significant similarity to proteins in the "nr" and CDD databases (Table 3). Notably, many of these proteins were rather short in length (e.g., *B. cinerea*) (Supplemental Fig. S11), an observation that may raise concerns about reliability of annotation of the corresponding genes.

Genes predicted but not annotated

We did not observe an excess of short genes in this category (e.g., *F. oxysporum*) (Supplemental Fig. S11). The fraction of the encoded proteins that exhibit similarity to proteins in the "nr" database was larger than in the group of genes annotated but not predicted (Table 4). Many of these similarities were detected with "nr" database proteins characterized as "hypothetical." Such findings support the ab initio predictions, though they do not shed light on the protein function. Still, similarities detected to known proteins led to several interesting findings as described in the Supplemental Results.

Comparison with other gene prediction programs

For a fair comparison of gene prediction programs using conventional supervised training, the programs have to be trained on the same data set that does not overlap with the test set. In assessment of a self-training program, the "no-overlap of training and test sets" condition can be fulfilled if the test set sequences are excluded from input to the self-training algorithm. However, a program that employs unsupervised training does not make any use of annotated genomic regions provided as a training set. Thus, a program with supervised training has a certain information advantage, unless the training set is biased (which is a frequent concern). Therefore, how to conduct a fair comparison of supervised and unsupervised programs is not immediately obvious. Thus, we conducted the program comparisons "as is" with an understanding that rigorous rules of fair competitions are yet to be determined.

The GipsyGene program is one of a few designed specifically for fungal genomes (Neverov et al. 2003). It was trained for *A. nidulans* and *N. crassa*. An addition of the BP model to the GipsyGene program led to increase of internal exon prediction Sn from 69%–75% to 75%–80% (with upper limits cited for *A. nidulans*). Given the difference in the test sets, it is reasonable to compare only Sn values, which in our case are 82.6%–83.6% for the original algorithm and 89.1%–91.8% for the algorithm with enhanced intron submodel (Supplemental Table S6). While a significant improvement in prediction accuracy was reached by GipsyGene after addition of the BP model, the overall accuracy remained relatively low, a fact apparently related to a small size of available training sets (193 and 99 genes for *A. nidulans* and *N. crassa*, respectively).

We also made a performance comparison with the locally run gene finder AUGUSTUS (Stanke and Waack 2003). For this purpose, we used the test sets of sequences of five fungal genomes with verified complete genes (Supplemental Table S4, Set I). In terms of average accuracy, (Sn + Sp)/2, the new program has shown better results in all eight categories for *F. verticulloides*, *M. grisea*, and *S. pombe* and in seven out of eight categories for *C.*

Table 3. Statistics of comparative genomics analysis of protein products of the genes annotated in the nine fungal genomes

Species	Total no. of annotated genes	No. of protein products with no similarity to predicted proteins	Proteins (from B) with similarity to proteins of other species in nr database	Percent of B	Proteins (from B) with conserved domains (similarity to CDD)	Percent of B
	A	B	C	D	E	F
A. niger	14,101	2851	107	3.75	26	0.91
B. cinerea	16,448	4413	218	4.94	43	0.97
C. immitis	10,457	2005	145	7.23	34	1.70
F. graminearum	13,332	1024	51	4.98	13	1.27
F. oxysporum	17,735	630	47	7.46	3	0.48
M. grisea	12,841	1491	39	2.62	10	0.67
R. oryzae	17,467	1086	17	1.57	10	0.92
S. sclerotiorum	14,522	3655	226	6.18	29	0.79
S. nodorum	16,597	2746	169	6.15	16	0.58

Table 4. Statistics of comparative genomics analysis of protein products of genes newly predicted in the nine fungal genomes

Species	Total no. of predicted genes	No. of predicted protein products with no similarity to predicted proteins	Proteins (from B) with similarity to proteins of other species in nr database	Percent of B	Proteins (from B) with conserved domains (similarity to CDD)	Percent of B
	A	B	C	D	E	F
A. niger	11,342	263	115	43.73	23	8.75
B. cinerea	11,890	529	153	28.92	40	7.56
C. immitis	8435	311	94	30.23	29	9.32
F. graminearum	12,426	288	121	42.01	27	9.38
F. oxysporum	20,843	1408	561	39.84	107	7.60
M. grisea	11,850	346	94	27.17	36	10.40
R. oryzae	16,477	446	108	24.22	71	15.92
S. sclerotiorum	11,119	342	112	32.75	21	6.14
S. nodorum	13,707	285	33	11.58	7	2.46

immitis and *C. cinereus* (Table 5). Note that insufficiently high performance of AUGUSTUS for *F. verticillioides* could be explained by the use of the model parameters derived for the genome of *F. graminearum* (and selected for AUGUSTUS from a set of pre-computed models). For *M. grisea*, the new algorithm shows 8.3% higher Sn and 5.3% higher Sp in internal exon prediction accuracy, while AUGUSTUS performs marginally better in Sp for detection of initiation and termination sites and for nucleotides. Nevertheless, all the average values, (Sn + Sp)/2, for *M. grisea* are higher for the new algorithm than for AUGUSTUS. There are several reasons to emphasize the results obtained for *S. pombe*. First, the size of *S. pombe* genome, 12 Mb, is close to the minimum size (10 Mb) required for unsupervised training (Lomsadze et al. 2005). Second, a large number of experimentally verified *S. pombe* genes are known; thus, the *S. pombe* genome is a good target for an algorithm using supervised training. Third, the test set contains about a half of all *S. pombe* genes with multiple exons; thus the chance of overlap between the AUGUSTUS training set and the current test set is high. Still, performance of the unsupervised algorithm on the *S. pombe* genome is better in all categories with respect to (Sn + Sp)/2.

We have further extended the comparative tests to the *S. pombe* artificial chromosomes. In gene splitting, both AUGUSTUS and the new algorithm show almost identical error rates (two to three gene splits per 1277 genes). In chromosomes with intergenic regions shorter than 1100 nt, AUGUSTUS predicts a larger number of gene fusions (Fig. 5) and makes more errors in exact gene prediction (Supplemental Fig. S9). AUGUSTUS shows local increases (spikes) in error rates for gene merging and a decrease in the accuracy of exact gene prediction for chromosomes with 750- to 950-nt-long intergenic regions. These effects could be related to the settings made in AUGUSTUS for modeling intergenic regions. On the other hand, AUGUSTUS makes a fewer number of false-positive predictions in intergenic regions longer than 4500 nt (Supplemental Fig. S8).

Discussion

It is natural to anticipate that the line of development of eukaryotic gene prediction algorithms would follow the path of prokaryotic gene finding. For prokaryotes, algorithms that focused almost entirely on the prediction stage of the gene identification problem (i.e., how to find genes given a training sequence) evolved into automatic self-training systems capable of adjusting to genome-specific properties in the process of estimating algorithm parameters from anonymous sequence (Audic and Claverie 1998; Hayes and Borodovsky 1998; Salzberg et al. 1998; Besemer et al. 2001; Larsen and Krogh 2003). A recent dramatic increase in the number of eukaryotic targets of genome sequencing has necessitated the development of unsupervised ab initio gene prediction algorithms for eukaryotes. We demonstrated earlier that an iterative ab initio self-training training is feasible for eukaryotic genomes (Lomsadze et al. 2005). Here we have further developed this approach to reach high accuracy in the application to genomes with specific features of gene organization, such as a large group of fungal genomes.

The introduction of the enhanced intron submodel significantly improved the prediction accuracy of the donor and acceptor sites as well as exons and introns (Table 2; Supplemental Table S6). The new algorithm predicted internal exons in 12 fungal genomes with Sn and Sp averages higher than 90% (Table 2; Supplemental Table S6). Since the new intron submodel increases the accuracy of exon detection, the accuracy of prediction of translation termination sites increases due to improved accuracy of detection of the last exon and its reading frame.

Fungal genomes analyzed in this article have been annotated by experts using both intrinsic and extrinsic methods of gene prediction (Supplemental Table S8). For instance, *C. neoformans* was annotated via the combined use of Exonerate (www.ebi.ac.uk/~guy/exonerate/), Twinscan (Korf et al. 2001), SNAP (Korf 2004), GeneZilla (www.genezilla.org/), and Glimmer.hmm (Majoros et al. 2003). We consider the public annotation of genes in *C. neoformans* genome as one of the best in our sample (Tenney et al. 2004), similar in accuracy to annotation of the well-studied *S. pombe* genome. It is quite satisfying to report that a single gene prediction algorithm applied to the genomes of *C. neoformans* and *S. pombe*, though significantly different in terms of number of introns per gene, produced predictions that match the established gene annotations quite well (Table 1). In general, comparison of the annotated and predicted genes in the 16 fungal genomes shows the tendency to have more annotated genes than predicted (Table 1). Interestingly, while the total numbers of predicted and annotated genes may differ significantly, the predicted and annotated average intron densities are close to each other in the majority of cases (Table 1).

A short intron length is likely to be an indication that the intron definition mechanism is in use, which detects an intron part of the pre-mRNA by spliceosome, rather than an exon part. In the fungal species studied here, we observed predominantly short introns (<100 nt). However, splicing by intron definition may not be necessarily associated with formation of strongly conserved BP sites as is revealed in the case of *R. oryzae*, a Zygo-

Table 5. Comparison of the performances of the GeneMark-ES-2 program and the AUGUSTUS program

	S. pombe				*C. immitis*				*F. verticillioides*				*M. grisea*				*C. cinereus*			
	AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2	
Internal exon																				
Sn	87.7		**88.2**		**82.9**		82.8		82.9		**85.6**		80.9		**89.2**		82.2		**85.0**	
Sp	88.4	*88.1*	**89.6**	***88.9***	84.4	*83.7*	**93.0**	***87.5***	79.8	*81.4*	**91.2**	***88.4***	86.4	*83.7*	**91.7**	***90.5***	86.5	*84.4*	**89.7**	***87.4***
Intron																				
Sn	90.2		**91.0**		82.9		**84.1**		85.2		**90.7**		79.3		**89.3**		82.7		**86.8**	
Sp	**93.3**	*91.8*	92.7	***91.9***	89.5	*86.2*	**91.3**	***86.8***	89.8	*87.5*	**94.3**	***92.5***	87.6	*83.5*	**90.5**	***89.9***	89.0	*85.9*	**90.3**	***88.6***
Donor																				
Sn	92.6		**93.1**		86.3		**87.0**		90.8		**92.2**		85.2		**92.1**		86.4		**89.6**	
Sp	**95.1**	*93.9*	**95.1**	***94.1***	90.8	*88.6*	**94.4**	***90.5***	90.8	*90.8*	**96.2**	***94.2***	92.7	*89.0*	**93.9**	***93.0***	91.5	*89.0*	**93.6**	***91.6***
Acceptor																				
Sn	91.2		**92.8**		85.1		**86.8**		87.3		**91.9**		82.1		**93.2**		84.9		**87.8**	
Sp	94.4	*92.8*	**94.6**	***93.7***	91.9	*88.5*	**94.7**	***89.7***	92.6	*90.0*	**96.2**	***94.1***	93.0	*87.6*	**94.6**	***93.9***	91.6	*88.3*	**92.1**	***90.0***
Exon																				
Sn	85.9		**88.0**		76.9		**79.7**		76.7		**85.3**		78.7		**88.0**		78.3		**81.2**	
Sp	88.8	*87.4*	**89.2**	***88.6***	83.0	*80.0*	**84.6**	***81.6***	80.3	*78.5*	**87.9**	***86.6***	88.5	*83.6*	**89.1**	***88.6***	**84.3**	*81.3*	84.3	***82.8***
Initiation site																				
Sn	83.9		**88.2**		74.3		**78.7**		70.6		**81.7**		76.9		**88.2**		70.7		**72.5**	
Sp	**90.1**	*87.0*	88.5	***88.4***	**87.9**	***81.1***	78.7	*79.0*	**84.0**	*77.3*	81.7	***81.7***	**94.9**	*85.9*	89.2	***88.7***	**81.4**	**76.1**	72.9	*72.7*
Termination site																				
Sn	91.5		**94.2**		79.6		**86.1**		82.3		**94.8**		76.9		**89.3**		78.4		**83.8**	
Sp	**96.0**	*93.8*	94.2	***94.2***	**91.5**	*85.6*	87.1	***86.6***	95.1	*88.7*	**95.4**	***95.1***	**92.9**	*84.9*	89.3	***89.3***	**89.7**	*84.1*	84.8	***84.3***
Nucleotide																				
Sn	96.2		**98.6**		90.4		**96.1**		95.9		**98.8**		87.8		**98.2**		90.9		**95.3**	
Sp	99.5	*97.9*	**99.6**	***99.1***	96.2	*93.3*	**96.5**	***96.3***	96.0	*96.0*	**97.1**	***98.0***	**96.4**	*92.1*	95.8	***97.0***	94.8	*92.9*	95.1	**95.2**

Values of Sn and Sp were determined for the test sets of complete genes (test sets of type I) (Supplemental Table S4). For gene prediction in *F. verticillioides*, the AUGUSTUS program uses model parameters derived in supervised mode for the *F. graminearum* genome. Boldface shows the larger value out of the two in corresponding category between AUGUSTUS and GeneMark-ES-2. Italicized values indicate average accuracy, (Sn + Sp)/2.

mycete. In contrast with the fungal species studied here from the phyla Ascomycota and Basidiomycota, *R. oryzae* short introns possess weak BP signals along with pronounced poly-Y tails near acceptor sites. Notably, fungal introns with strong BP sites, such as observed in the genomes of *C. immitis*, *M. grisea*, and *N. crassa* have even longer introns than *R. oryzae* (Supplemental Fig. S12). Interestingly, the exon length distributions of the species studied here show less variability than seen for the introns (Supplemental Fig. S13).

Conclusions

We anticipate that the ab initio self-training algorithm with enhanced intron submodel will be a useful tool for eukaryotic genomes with gene organizations similar to fungal genomes analyzed here. We have to emphasize that even within the fungal phyla, the deeply branched lineages are ubiquitous and the diversity is significant. Nevertheless, the algorithm adaptive training strategy allows for the automatic adjustment of the HMM architecture and the corresponding parameters. Probabilities of transitions between hidden states in the enhanced intron submodel are tuned to fit genomes with various numbers of introns per gene, as well as genomes with different types of splicing mechanisms: those that use information either from the BP sites or from the poly-Y tails. We have demonstrated that prior knowledge of the type of the splicing mechanism is not necessary for the application of the algorithm. Moreover, the algorithm is able to adapt to a genome with an a priori unknown fraction of introns possessing conserved BP sites. In our tests, we observed that the algorithm predicted splice sites in 13 fungal genomes with higher than 90% accuracy (Sn and Sp).

The advantage of the new ab initio gene prediction tool is twofold. First, it shows high performance; second, it is user friendly for the analysis of novel genomes, since it excludes an elaborate development of a training set. We also have to say that the self-training procedure is amenable to incorporation of extrinsic data (work in progress). We anticipate that the self-training approach will facilitate simplification and standardization of the architecture of gene annotation pipelines. While self-training seems to be especially valuable for early stages of sequencing projects, even at more advanced stages of the project with ample extrinsic information available, the self-training algorithm provides an integration framework for accurate and consistent gene annotation.

Acknowledgments

We thank Andrey Kislyuk for help in developing the software program used for preparation of the test sets; Marc Bruce, Igor Grigoriev, and Chinnappa Kodira for useful discussions; King Jordan for useful comments on the manuscript; and Wenhan Zhu and Ryan Mills for help with manuscript preparation. This work was supported in part by grant GM58763 to Y.O.C. and grant HG00783 to M.B. from the National Institutes of Health (NIH). Funding to pay publication charges for this article was provided by NIH grants GM58763 and HG00783.

References

Audic, S. and Claverie, J.M. 1998. Self-identification of protein-coding regions in microbial genomes. *Proc. Natl. Acad. Sci.* **95:** 10026–10031.

Besemer, J. and Borodovsky, M. 1999. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* **27:** 3911–3920.

Besemer, J., Lomsadze, A., and Borodovsky, M. 2001. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29:** 2607–2618.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and GenomeWise. *Genome Res.* **14:** 988–995.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Comput. Biol.* **268:** 78–94.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23:** 673–679.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A. 1996. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci.* **93:** 9061–9066.

Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., et al. 2006. EGASP: The human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7:** 1–31.

Hayes, W.S. and Borodovsky, M. 1998. How to interpret an anonymous bacterial genome: Machine learning approach to gene identification. *Genome Res.* **8:** 1154–1171.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., and Brunak, S. 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24:** 3439–3452.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59. doi: 10.1186/1471-2105-5-59.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** 140–148.

Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 179–186.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4:** 134–142.

Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., and Murphy, J.W. 2004. Introns and splicing elements of five diverse fungi. *Eukaryot. Cell* **3:** 1088–1100.

Larsen, T. and Krogh, A. 2003. EasyGene—A prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4:** 21. doi: 10.1186/1471-2105-4-21.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262:** 208–214.

Lim, L.P. and Burge, C.B. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci.* **98:** 11193–11198.

Liu, W.S. and Ponce de Léon, F.A. 2004. Assignment of SRY, ANT3, and CSF2RA to the bovine Y chromosome by FISH and RH mapping. *Animal Biotechnol.* **15:** 103–109.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y., and Borodovsky, M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33:** 6494–6506.

Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.* **26:** 1107–1115.

Lukashin, A.V., Engelbrecht, J., and Brunak, S. 1992. Multiple alignment using simulated annealing: Branch point definition in human mRNA splicing. *Nucleic Acids Res.* **20:** 2511–2516.

Majoros, W.H., Pertea, M., Antonescu, C., and Salzberg, S.L. 2003. GlimmerM, Exonomy and Unveil: Three ab initio eukaryotic gene finders. *Nucleic Acids Res.* **31:** 3601–3604.

Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., et al. 2007. CDD: A conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* **35:** 237–240.

Mathe, C., Sagot, M.F., Schiex, T., and Rouze, P. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30:** 4103–4117.

Mitrophanov, A.Y., Lomsadze, A., and Borodovsky, M. 2005. Sensitivity of hidden Markov models. *J. Appl. Probab.* **42:** 632–642.

Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477–478.

Neverov, A., Gelfand, M., and Mironov, A. 2003. GipsyGene: A statistics-based gene recognizer for fungal genomes. *Biophysics* **48:** S71–S75.

Parra, G., Blanco, E., and Guigo, R. 2000. GeneID in *Drosophila*. *Genome Res.* **10:** 511–515.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., and Rouze, P. 1999. Evaluation of gene prediction software using a genomic data set: Application to *Arabidopsis thaliana* sequences. *Bioinformatics* **15:** 887–899.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10:** 483–501.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26:** 544–548.

Stanke, M. and Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19:** 215–225.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24:** 637–644.

Tenney, A.E., Brown, R.H., Vaske, C., Lodge, J.K., Doering, T.L., and Brent, M.R. 2004. Gene prediction and verification in a compact genome with numerous small introns. *Genome Res.* **14:** 2330.

Thompson, W., Rouchka, E.C., and Lawrence, C.E. 2003. Gibbs Recursive Sampler: Finding transcription factor binding sites. *Nucleic Acids Res.* **31:** 3580–3585.

Watanabe, K. and Harayama, S. 2001. SWISS-PROT: The curated protein sequence database on Internet. *Tanpakushitsu Kakusan Koso* **46:** 80–86.

Received May 31, 2008; accepted in revised form August 26, 2008.